DOCUMENT RESUME

ED 462 436 TM 033 700

AUTHOR Felan, George D.
TITLE Test Equating: Mean, Linear, Equipercentile, and Item Response Theory.
PUB DATE 2002-02-16
NOTE 24p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, February 14-16, 2002).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Equated Scores; *Item Response Theory
IDENTIFIERS *Equipercentile Equating; *Linear Equating Method

ABSTRACT

This paper discusses the four major types of test equating: (1) mean; (2) linear; (3) equipercentile; and (4) item response theory. The single-group, equivalent-group, and anchor-test data collection designs are presented as methods used for test equating. Issues related to assumptions and equating error are also addressed. The advantages and disadvantages of each equating method are discussed along with the conditions conducive to satisfactory equating. Research on the current interest in equating state tests to the Voluntary National Test or the National Assessment of Educational Progress is reviewed, and it is concluded that such equating is unlikely. (Contains 21 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 462 436

Running head: TEST EQUATING

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

G. Felan

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Test Equating: Mean, Linear, Equipercentile, and Item Response Theory

George D. Felan

University of North Texas

TM033700

Correspondence should be addressed to George D. Felan, 3507 N. Main St., Fort Worth, TX 76106-4348, or E-Mail: georgedfelan@hotmail.com

Paper presented at the annual meeting of the Southwest Educational Research Association, February 16, 2002, Austin, TX.

BEST COPY AVAILABLE

Abstract

This article will discuss the four major types of test equating: (a) mean, (b) linear, (c) equipercentile, and (d) item response theory. The single-group, equivalent-group, and anchor-test data collection designs will also be presented as methods used for test equating. Issues related to assumptions and equating error are also addressed. The advantages and disadvantages of each equating method are the discussed along with the conditions conducive to satisfactory equating. Finally, research on the current trend to equate state tests to the Voluntary National Test or the National Assessment of Educational Progress is reviewed with the conclusion drawn that such equating is highly unlikely.

Test equating

Test equating is a statistical procedure to establish the relationships between scores from two or more tests. This procedure is also used to place two or more tests on a common scale. Loosely termed, test equating may be referred to as linking, calibration, and scaling (Kolen & Brennan, 1995). Test equating is often used in situations where multiple forms of a test exist, and examinees taking different forms are compared to each other or if researchers want to overcome problems of practice effects. In addition to statistical procedures, successful equating involves many aspects of testing, including procedures to develop tests, to administer and score tests, and to interpret scores earned on tests.

In the early 1980s, the importance of equating began to be recognized by a broader spectrum of people associated with testing (Woldbeck, 1998). For example, the American Educational Research Association (AERA), American Psychological Association (APA), and the National Commission for Measurement in Education (NCME) (1999*) Standards for Educational and Psychological Testing* devoted a substantial portion of a chapter to equating, whereas the previous edition did not even list equating in the index (Kolen & Brennan, 1995).

The prominence of equating, along with its interdependence with so many aspects of the testing process, also suggests that test developers and all other testing professionals should be familiar with the concepts, statistical procedures, and practical issues associated with equating. Experience suggests that relatively few measurement professionals have sufficient knowledge to conduct test equating (Kolen & Brennan, 1995). Also, many do not fully appreciate the practical consequences of various changes in testing procedures on equating, such as the consequences of many test-legislation initiatives, the use of performance assessments, and the introduction of computerized test administration.

Test equating is needed in order to improve test score integrity by not having to administer the same test again and again. Test equating is also used as a way to ensure fairness of a test or eliminate practice effect. Therefore, multiple forms often are required in testing practice (Kolen & Brennan, 1995).

Test equating is the preferred method when there is an issue of test score exchangeability. Furthermore, while multiple tests are being used to measure the same variable in practice, test scores from these tests are often exchangeable because they are set on different scales. The last reason is for test continuity (vertical) which allows for multiple tests being used at different ability levels to measure growth or change of an ability or trait (Kolen & Brennan, 1995).

There are two groups of item statistics used in conducting item analyses (Zhu, 1998). The first group, commonly called classical item statistics, includes an "item difficulty" statistic (proportion who answer the item correctly or p-value) and an "item discrimination" statistic (point-biseral correlation). These statistics are population dependent, that is, they can vary from examinee group to examinee group due to the knowledge and skill level of the group of examinees who challenged the items (Woldbeck, 1998). If the examinees are "smart", the item appears to be easy. If the examinees are less knowledgeable, the item appears to be more difficult because fewer people answer it correctly.

The second group of item statistics is generated from the Rasch model analysis and includes an "item difficulty calibration" statistic, a "calibration error" statistic and an "item fit" statistic (Kolen & Brennan, 1995). The item difficulty calibration statistic, estimates the location or relative difficulty of the item on the equal interval logit scale using a log-linear formula. The calibration error statistic documents the measurement error associated with the item difficulty

calibration. The item fit statistics estimate how closely the item follows the expectations of the Rasch model (Kolen & Brennan, 1995), namely that those who are more knowledgeable will answer the item correctly more frequently than those who are less knowledgeable.

Horizontal and vertical test equating

There are two types of test equating, horizontal and vertical. Horizontal equating is appropriate when multiple forms of a test are required to maintain test security. The forms are expected to be parallel in content and difficulty (Kolen, 1988). Furthermore, equating procedures do not function well when there are large differences in form-to-form difficulty, reliability, or test content. Ability distributions of examinees are expected to be approximately equal. If not, traditional equating methods (e.g., linear and equipercentile equating) may not be appropriate (Kolen, 1988).

Vertical methods equate scores on two tests intentionally designed to be different in difficulty but still measure the same general knowledge or domain or skills. Unlike horizontal equating, the ability distribution of examinees at the various levels will be different. Barnard (1996) rightly points out that in vertical equating the contents of forms at different levels is different and therefore the scores on such tests cannot be used interchangeably. Since equating adjusts for difficulty differences and not for differences in content, this dimensionality problem leads to the use of the term vertical scaling rather than vertical equating. The term equating is therefore, in accordance with the suggestion in the *Standards for Educational and Psychological Testing* (American Psychological Association, 1999) reserved for the process when the transformations are made between forms of comparable difficulty which measure the same underlying construct or contents (Kolen & Brennan, 1995).

The process of equating should not be confused with the process of scaling. The

difference can best be explained by means of an example. Suppose that a testee answers 30 of the 40 questions of one form of a test correctly and assume that each question counts one mark. This raw score is converted to another convenient scale (such as percentages). This process is referred to as scaling. This conversion can be done by linear or non-linear transformations (Barnard, 1996). The main advantage of the former is its simplicity while the latter is used for reasons of flexibility. The new scale is called the primary score scale. The testee's raw score of 30 will thus be 75 on the primary score scale. This is not equating, only scaling (Han, 1997).

Now assume that, according to an equating process, it was found that a second form of the test is uniformly four score points easier than the first form. A raw score of 34 on the second form will thus be equal to a raw score of 30 on the first form which is equal to a score of 75 on the primary score scale. The scores on the primary score scale means the same, irrespective of the form from which they were derived. Reported scores are all on the same scale and can be used interchangeably. The equating process has thereby adjusted for the difference in difficulty since primary score scales have the same meaning regardless of the form a test takes.

Equating assumptions

The tests to be equated must measure the same ability (or characteristics, traits, or skills). Equity means that the conditional frequency distribution of scores on Test A, after equating, is the same as the distribution of scores on Test B. Scores therefore should be interchangeable after equating (Zhu, 1998). Threats to this assumption include forms of a test that are either too easy or too hard and thus produce score distributions that do not reflect the distribution of scores from the previous test form (this is also a threat to reliability).

Population invariance means that the test equating should be independent of the data or examines employed in the equating process, and a conversion derived from the equating should apply to all similar situations (Kolen, 1988).

Symmetry means that the transformation should be the same regardless of which test is used as the converting reference or base (i.e., interpretation of test scores should be the same based on either the equating from Test 1 to Test 2 or that from Test 2 to Test 1) (Kolen, 1988).

Data collection designs

The three commonly used data collection designs are (a)single-group, (b)equivalent-groups, and (c) anchor-test designs. In the single-group design, two or more testing forms are administered to the same group of examinees. Measurement error is relatively small since there is only one group of examinees. The major factors to be concerned with are fatigue and practice effects. To avoid fatigue and practice effects, either a spiraling process should be applied or the order of testing forms can be counterbalanced.

The equivalent-groups design (random-groups design) method involves two tests administered to two equivalent groups of examinees. The advantages in using this method include the fact that fatigue and practice effects are eliminated and testing time is minimized. A negative factor is the unknown degree of bias introduced because groups often are not the same in terms of their ability distributions. To control for this bias, larger groups are generally required for this design (Zhu, 1998).

The third method commonly used is the anchor-test design (common-item nonequivalent groups design). This procedure requires the tests to be administered to two different groups of examinees. The groups can be different from each other in their ability distributions. A set of common or anchor items is included in both tests or forms. Differences between the two tests

can be adjusted based on common-item statistics (Zhu, 1998). This procedure is useful in measuring growth when two groups are known to be non-equivalent. This is also useful when it is impossible to administer more than one test per test date due to test security or other practical concerns (Woldbeck, 1998). Anchor-test design is also necessary when developing an item bank, in which testing items are cumulated into a common scale. The use of this procedure requires strong statistical assumptions for effects of group and test differences, therefore, there should be enough common items with representative content to be measured. A rule of thumb for the minimum length of an anchor test is 20-25% of the number of items on either of the tests (Woldbeck, 1998).

Traditional equating methods

The two types of equating methods are (a) traditional equating methods, and (b) item response theory equating methods. Traditional equating methods are based on Classical Testing Theory (CTT) whereby observed scores are believed to consist of "true scores" and "errors." I will first address the traditional equating methods and then I will address the issues involved in item response theory equating. The three types of traditional equating methods are (a) mean equating, (b) linear equating, and (c) equipercentile equating (Barnard, 1996).

In mean equating, the means of two forms are set equal for a particular group of examinees; that is, the Form 2 scores are converted so that their mean will equal the mean of the scores on Form 1. This type of equating assumes difference in difficulty between the forms is constant throughout the entire score range (Barnard, 1996).

The second type of traditional equating is known as linear equating. In linear equating, the means and standard deviations on the two forms for a particular group of examinees are set equal; that is, Form 2 scores are converted so as to have the same mean and standard deviation as

scores on Form 1. This type of equating allows the relative difficulty of the forms to vary along the score scale. For example, Form 1 might be relatively more difficult than Form 2 for low achieving students than for high achieving students (Harris & Kolen, 1990).

The final type of traditional equating is equipercentile equating. In this type of equating, the Form 2 distribution is set equal to the Form 1 distribution for a particular group of examinees by scoring the two tests as percentages. Form 2 scores that are converted using equipercentile equating have approximately the same mean, standard deviation, and distributional shape (skewness, kurtosis, etc.) as do scores on Form 1. This provides for even greater similarity between distributions of equated scores than does linear equating. Scores on Form 1 and Form 2 with the same percentile rank for a particular group of examinees are considered to indicate the same level of performance (Zhu, 1998). At this point, it would be appropriate to refer to Appendix A and Appendix B. Note the comparisons between the (a) mean, (b) standard deviation, (c) alpha, (d) standard error of measurement (SEM), and (e) mean biserial.

The following guidelines should be followed in choosing between the different methods of equating. If forms to be equated have equal standard deviations, then mean equating and linear equating will produce the same results. If the distributions have the same shape (skewness, kurtosis, etc.), then the linear and equipercentile methods produce the same results. Equipercentile equating typically requires larger sample sizes than does linear or mean equating and is more complex computationally (Zhu, 1998).

Item response theory equating

Item response theory equating is also known as latent trait theory or item characteristic curve theory. This theory represents a mathematical model describing how examinees at different ability levels should respond to an item for the trait to be measured (Cook & Eignor,

1991). The process of item response theory equating begins with data collection. The anchor-test design is the most commonly used because group abilities often differ from each other and traditional equating methods often will not work well in this circumstance (Cook & Eignor, 1991).

The next step is to select an appropriate item response theory model. The commonly used item response theory models for dichotomous scores are the one-(Rasch), two-, and three-parameter logistic models (Woldbeck, 1998). With the selected item response theory model item and examinee parameters can then be estimated, which is usually accomplished by employing certain computer programs. Model-data fit is also examined statistically at the same time. If the model and data do not fit, either a new model should be considered or new data should be collected (Woldbeck, 1998).

If the model and data fit, the equating can then move to the next step, in which parameter estimates from separate calibrations are placed on a common scale. Scaling constants can generally be classified into four categories (Kolen & Brennan, 1995): (a) regression, (b) mean and sigma, (c) robust means and sigma, and (d) characteristic curve methods. Item response theory equating has basically been completed after parameters from two separate estimations have been set on the same scale.

Advantages of item response theory equating

A main advantage of IRT equating is that item parameters are independent of the ability level of examinees responding to the item, and at the same time, ability is also independent of the performance of other examinees and the items used in tests (Zhu, 1998). This is known as the "invariance" feature of item response theory. Therefore, the interpretation of item difficulty and

examinee ability is consistent in item response theory. The invariance feature also decreases the impact of sampling error (Zhu, 1998).

Secondly, the precision of measurement can be determined at any ability level. In classical test theory, this is usually determined at the group level. Also, item difficulty and examinee ability are set on the same scale, which makes it much easier to determine the appropriateness of item difficulty and interpret test scores (Woldbeck, 1998). Item response theory equating therefore produces better equating at the upper end of the score scale. Greater flexibility is also provided in choosing previous test versions. This is because if an item is dropped, the shortened test can be easily reconstructed based on the item information from the remaining test items. In the final analysis, preequating becomes possible even before a test is administered.

A caveat that must be noted is that item response theory equating is not always superior to traditional equating. A number of researchers have compared the performance of these equating methods and found that traditional equating often worked as well as item response theory equating methods, with the possible exception of the anchor-test design with nonequivalent groups (Harris & Kolen, 1990).

Equating error

Equating error may occur as random or systematic. Random equating error presents itself whenever samples from populations of examinees are used to estimate parameters such as means, standard deviations, and percentile ranks (Barnard, 1996). This type of error can be reduced by using larger samples and by choice of equating design.

The other type of error that can occur is systematic equating error. Systematic equating error results from violations of the assumptions and conditions of the particular equating

methodology used (Zeng, 1991). In the single group design, failure to control for fatigue and practice effects can be a major source of systematic error. In the random groups design, systematic error will result if the spiraling process is ineffective in achieving group comparability. Systematic error is especially problematic in the nonequivalent groups design and will result if the assumptions underlying the method used are not met (Zeng, 1991). Assumptions can be especially difficult to meet if the groups differ substantially, or if the common items are not representative of the total test form in content and statistical characteristics. Furthermore, systematic error will likely result if the common items function differently from one administration to another. This can occur if their position on the old and new form is not the same. For any equating design, systematic error can also result if the new form and old form differ in content, difficulty, and reliability (Zeng, 1991).

One way to control for random or systematic equating error is through the use of an adequate sample size. When using linear equating, a sample size of 400 per form is usually preferred (Zeng, 1991). When using equipercentile equating, a sample size of 1500 per form or test is preferred. When using item response theory equating, a sample size of 400 for the Rasch model or 1500 for the three-parameter model is preferred. If sample size is small, using log-linear smoothing or the collateral information method might help overcome sample size problems (Zeng, 1991).

Conditions conducive to satisfactory equating

In general, the goals of equating, such as equating accuracy and the extent to which scores are to be comparable over long time periods, are to be clearly specified. The design for data collection, the equating linkage plan, the statistical methods used, and the procedures for choosing among results, should be appropriate for achieving the goals in the particular practical

context in which equating is conducted. Finally, adequate quality control procedures must be followed (Kolen & Brennan, 1995).

In terms of test development using all designs, test content and statistical specifications should be well defined and stable over time. When the test form is constructed, statistics on all or most of the items should be available from pretesting or previous use. The test should be reasonably long (e.g., 30 items or longer). Scoring keys should be stable when items or forms are used on multiple occasions (Kolen & Brennan, 1995).

In terms of test development, when using common-item nonequivalent groups design, each common set should be representative of the total test in content and statistical characteristics. Each common-item set should be of sufficient length (e.g., at least 20% of the test for tests of 40 items or more; at least 30 items for long tests). Each common-item should be in approximately the same position in the old and new forms. Common-item stems, alternatives, and stimulus materials (if applicable) should be identical in the old and new forms. Other item level context effects should be controlled. If double linking is used, one old form should be administered during the same time of year as the form to be equated and one old form should be administered within a year or so (Kolen & Brennan, 1995).

In terms of examinee groups, they should be (a) representative of operationally tested examinees, (b) stable over time, (c)relatively large, and (d) in the common-item nonequivalent groups design. The groups taking the old and new forms should not be extremely different (Kolen & Brennan, 1995).

In terms of administration, the test and test items should be secure and administered under carefully controlled standardized conditions that are the same each time the test is

administered (Kolen & Brennan, 1995). Finally, in terms of field of study, the curriculum, training materials, and/or field of study should be stable (Kolen & Brennan, 1995).

Issues in test equating

The increased attention to test equating has been furthered by an expansion in the number of testing programs that use multiple forms which have to be equated. Also, test developers have referenced the role of equating in arriving at reported scores to address issues raised by testing critics while the accountability movement in education and resultant issues of fairness in testing have become much more visible (Kolen & Brennan, 1995).

Consider the fact that our own president is calling for a National Voluntary Testing program as a form of nationalized accountability testing. How will state assessments such as the Texas Assessment of Academic Skills (TAAS) measure up to such a national testing program? The following research studies have compared certain state tests to the National Assessment of Academic Progress (NAEP) and have discovered that these tests can't be equated.

The content of a test is shaped by the kinds of knowledge and skills addressed in its questions. Content is not generally comparable among various state assessments and commercial tests, even when they are testing the same subjects. Middle-school mathematics, for instance, covers several subject areas of knowledge, such as arithmetic, algebra, and geometry. The content of one state's 8th grade mathematics test might focus largely on multiplication, division, and other number operations skills, while another test may stress pattern recognition and other pre-algebra skills (Bond & Jaeger, 1993). In reading, one 4th grade test may emphasize vocabulary and basic comprehension, while another may give greater weight to critical evaluation of an author's themes (Afflerbach, 1995).

A related content issue pertains to the skills and cognitive processes required to answer items. Off-the-shelf commercial tests and tests that are custom developed for states are increasingly constructed as mixed-model assessments that contain different types of items, including multiple-choice items and various kinds of open-ended questions for which students construct their own responses by filling in a blank, solving a problem, writing a short answer, writing a longer response, or completing a graph or diagram (Shavelson, Baxter, & Pine, 1992). Colorado, Connecticut, North Carolina, and Maryland are examples of states with mixed-model assessments. Some item types are very useful for testing student recall of factual material (a claim often made for certain types of multiple choice items); other item types are better suited to eliciting direct evidence of how well a student can solve problems.

The effect of format differences on linkages can be substantial. For example, the 1991 National Assessment of Educational Progress trial state assessment in mathematics contained both multiple choice and short-answer formats. Linn, Shepard, and Hartka (1992) found that when the two formats were scored separately, there was enough difference between the scores to change the rank order of the states in the mathematics assessment. For items with constructed responses (that is, not multiple choice), variations in scoring may also influence the validity of linkages because different scoring guides may credit different aspects of performance, even when the items appear similar (Linn, 1993). Issues such as how the scorers are trained and which scoring guidelines they use can affect the objectivity and consistency of scoring (Frederiksen & Collins, 1989). Some states, including Vermont and New Mexico, are trying out new assessment formats, such as systematically evaluating collections ("portfolios") of a student's work, that raise even more complex issues about comparability and scoring (Valencia & Au, 1997).

In short, content, format, and related issues are vitally important in linking, and existing commercially developed achievement tests and state assessments differ substantially among themselves and National Assessment of Educational Progress on these dimensions. The lack of strong comparability in these areas prevents the development of reliable and valid linkages. In addition, statistical linkages between tests with substantial differences in content and degrees of difficulty will not be accurate in the sense that they will not be consistent across subpopulations. This lack of consistency is directly due to the differences in content and test difficulty.

In a study linking scores from the NAEP to statewide test results, Ercikan (1997) noted that results based on an equipercentile procedure suggest that such a link does not provide precise information.

RAND conducted several analyses to examine the issue of whether TAAS scores can be trusted to provide an accurate index of student skills and abilities. First, RAND used scores on the reading and math tests that are administered as part of the National Assessment of Educational Progress (NAEP) to investigate how much students in Texas have improved and whether this improvement is consistent with what has occurred nationwide. NAEP scores are a good benchmark for this purpose because they reflect national content standards and they are not subject to the same external pressures to boost scores as there are on the TAAS (Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. , 2000).

RAND's findings raise serious questions about the validity of the gains in TAAS scores. More generally, results illustrate the danger of relying on statewide test scores as the sole measure of student achievement when these scores are used to make high-stakes decisions about teachers and schools as well as students. It is anticipated that the findings will be of interest to

local, state, and national educational policymakers, legislators, educators, and fellow researchers and measurement specialists (Klein, et al., 2000).

In a study by Linn and Kiplinger (1995), the adequacy of linking statewide standardized test results to the NAEP by using equipercentile equating procedures was investigated using statewide mathematics data from four states. The results suggested that the linkings are not sufficiently trustworthy to make comparisons based on the tails of the distribution.

A study was conducted by Feuer, Holland, Green, Bertenthal, and Hemphill (1999) of the feasibility of establishing an equivalency scale that would enable commercial state tests to be linked to one another and to the NAEP. In evaluating the feasibility of linkages, the study committee focused on the linkage of various forth-grade reading tests and the liknage of various eighth-grade mathematics tests. Committee members concentrated on the factors that affect the validity of the inferences about student performance that users would draw from the linked test scores. The committee concluded that comparing the full array of currently administered commercial and state achievement tests to one another, through the development of a single equivalency or linking scale, is not feasible. Nor is reporting individual student scores from the full array of tests on the NAEP scale and transforming individual score on these tests and assessments into NAEP achievement levels feasible. In conclusion, unless the test to be linked to the NAEP is very similar in content, format, and uses, the resulting linkage is likely to be unstable and potentially misleading.

Cizek (2000) looked at issues involved in linking the NAEP to the proposed Voluntary National Tests (VNT). He noted that there are substantial differences between NAEP and the VNT which present serious challenges to linking the VNT to the NAEP. The single greatest consideration in evaluating the potential for score scales to be linked is that of construct

equivalence. Furthermore, Cizek (2000) notes that the greatest impact on overall construct equivalence is the extent to which content covered on the proposed VNTs can be viewed as consistent with that covered by the respective NAEP tests. Achievement levels, reporting methods, interpretations, and audiences must also be considered. The technical problems are serious enough, and the weight of policy considerations and uncertainty about how a VNT will affect NAEP are also worth contemplating. Cizek (2000) concludes there are policy issues that should be addressed before considering linking methods.

References

Afflerbach, P. A. (1995). *Content validation of the 1994 National Assessment of Educational Progress in reading: Classifying items according to the reading framework.* Stanford, CA: The National Academy of Education.

Barnard, J. J. (1996). *In search for equity in educational measurement: Traditional versus modern equating methods.* Paper presented at ASEESA's national conference at the HSRC Conference Centre, Pretoria, South Africa.

Bond, L., & Jaeger, R. M. (1993*). Judged congruence between various state assessment tests in mathematics and the 1990 National Assessment of Educational Progress item pool for grade-8 mathematics.* Greensboro, NC: University of North Carolina Center for Educational Research and Evaluation

Cizek, G. J. (2000*). Factors affecting linkage of the National Assessment of Educational Progress and the proposed Voluntary National Test.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA, (ERIC Document Reproduction Service No. ED 447 196).

Cook, L. L., & Eignor, D. R. (1991). *An NCME instructional module on item response theory equating methods.* Educational Measurement: Issues and Practices, 10, 37-45.

Ercikan, K. (1997*). Linking statewide tests to the National Assessment of Educational Progress: Accuracy of combining test results across states.* Applied Measurement in Education, 10 (2), 145-159.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: equivalence and linkage among educational tests.* (ERIC Document Reproduction Service No. ED 440 984).

Frederiksen, J. R., & Collins, A. (1989). *A systems approach to educational testing.* Educational Researcher, 18 (9), 27-32.

Han, T. (1997). *A comparison among item response theory true and observed score equatings and traditional equipercentile equating.* Applied Measurement in Education,10 (2), 105-121.

Harris, D., & Kolen, M. J. (1990). *A comparison of two equipercentile equating methods for common item equating.* Educational and Psychological Measurement,50 (1), 61-71. (ERIC Document Reproduction Service No. EJ 407 937).

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* [on-line]. Available: http://www.rand.org/publications/IP/IP202/

Kolen, M. J., & Brennan, R. L. (1995). Test equating: Methods and practices. New York: Springer.

Kolen, M. J., (1988). *An NCME instructional module on traditional equating methodology.* Educational Measurement: Issues and Practices, 7, 29-36.

Linn, R. L. (1993). *Linking results of distinct assessments.* Applied Measurement in Education, 6 (1), 83-102.

Linn, R. L., & Kiplinger, V. L. (1995). *Linking statewide tests to the National Assessment of Educational Progress: Stability of results.* Applied Measurement in Education, 8 (2), 135-155.

Linn, R. L., Shepard, L., & Hartka, E. (1992). *The relative standing of states in the 1990 trial state assessment: The influence of choice of content, statistics, and subpopulation breakdowns in Studies for the Evaluation of the National Assessment of Educational Progress Trial State Assessment.* Stanford, CA: The National Academy of Education.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). *Performance assessments: Political rhetoric and measurement reality.* Educational Measurement, 21 (4), 22-27.

Valencia, S. W., & Au, K. H. (1997). *Portfolios across educational contests: Issues of evaluation, teacher development, and system validity.* Educational Assessment, 4 (1), 1-35.

Woldbeck, T. (1998). *Basic concepts in modern methods of test equating.* Paper presented at the Annual Meeting of the Southwest Psychological Association, New Orleans, LA, (ERIC Document Reproduction Service No. ED 417 215).

Zeng, L. (1991). *Standard errors of linear equating for the single-group design.* ACT Research Report Series. 91-4.

Zhu, W. (1998). Test equating: what, why, how? Research Quarterly for Exercise and Sport, 69, 11-23.

Appendix A

ITEMAN (tm) for 32-bit Windows, Version 3.6 Page 6
Copyright (c) 1982 - 1998 by Assessment Systems Corporation

Conventional Item and Test Analysis Program

Item analysis for data from file C:\My Documents\Form X1.txt
Date: Dec 13, 2001 Time: 12:09 AM

There were 1655 examinees in the data file.

Scale Statistics

Scale: 0

N of Items	36
N of Examinees	1655
Mean	15.821
Variance	42.612
Std. Dev.	6.528
Skew	0.580
Kurtosis	-0.278
Minimum	2.000
Maximum	36.000
Median	15.000
Alpha	0.842
SEM	2.594
Mean P	0.439
Mean Item-Tot.	0.387
Mean Biserial	0.507
Max Score (Low)	11
N (Low Group)	483
Min Score (High)	20
N (High Group)	460

Appendix B

ITEMAN (tm) for 32-bit Windows, Version 3.6 Page 6
Copyright (c) 1982 - 1998 by Assessment Systems Corporation

Conventional Item and Test Analysis Program

Item analysis for data from file C:\My Documents\Form Y1.txt
Date: Dec 13, 2001 Time: 12:12 AM

There were 1638 examinees in the data file.

Scale Statistics

Scale: 0	
N of Items	36
N of Examinees	1638
Mean	18.673
Variance	47.313
Std. Dev.	6.878
Skew	0.205
Kurtosis	-0.697
Minimum	3.000
Maximum	36.000
Median	18.000
Alpha	0.860
SEM	2.577
Mean P	0.519
Mean Item-Tot.	0.409
Mean Biserial	0.536
Max Score (Low)	14
N (Low Group)	495
Min Score (High)	23
N (High Group)	499